Chapter 11
A Classification Method Based on Principal Component Analysis and Differential Evolution Algorithm Applied for Prediction Diagnosis from Clinical EMR Heart Data Sets

Pasi Luukka and Jouni Lampinen

Abstract. In this article we have studied the usage of a classification method based on preprocessing the data first using principal component analysis, and then using the compressed data in actual classification process which is based on differential evolution algorithm, an evolutionary optimization algorithm. This method is applied here for prediction diagnosis from clinical data sets with chief complaint of chest pain using classical Electronic Medical Record (EMR), heart data sets. For experimentation we used a set of five frequently applied benchmark data sets including Cleveland, Hungarian, Long Beach, Switzerland and Statlog data sets. These data sets are containing demographic properties, clinical symptoms, clinical findings, laboratory test results specific electrocardiography (ECG), results pertaining to angina and coronary infarction, etc. In other words, classical EMR data pertaining to the evaluation of a chest pain patient and ruling out angina and/or Coronary Artery Disease, (CAD). The prediction diagnosis results with the proposed classification approach were found promisingly accurate. For example, the Switzerland data set was classified with 94.5% ± 0.4% accuracy. Combining all these data sets resulted in the classification accuracy of 82.0% ± 0.5%. We compared the results of the proposed method with the corresponding results of the other methods reported in the literature that have demonstrated relatively high classification performance in solving this problem. Depending on the case, the results of the proposed method were of equal level with the best compared methods, or outperformed their
classification accuracy clearly. In general, the results are suggesting that the proposed method has potential in this task.

11.1 Introduction

Many data sets that come from the real world are admittedly coupled with noise. Noise can be stated as a random error or variance of a measured variable [13]. Data analysis is almost always burdened with uncertainty of different kinds. There are several different techniques to deal with noisy data [7].

A major problem in mining scientific data sets is that the data is often high dimensional. In many cases there is a large number of features representing the object. One problem is that the computational time for the pattern recognition algorithms can become prohibitive, when the number of dimensions grows high. This can be a severe problem, especially as some of the features are not discriminatory. Besides the computational cost, irrelevant features may also cause a reduction in the accuracy of some algorithms.

To address this problem of high dimensionality, a common approach is to identify the most important features associated with an object, so that further processing can be simplified without compromising the quality of the final results. There are several different ways in which the dimension of a problem can be reduced. The simplest approach is to identify important attributes based on the input from domain experts. Another commonly used approach is Principal Component Analysis (PCA) [19], which defines new attributes (principal components or PCs) as mutually-orthogonal linear combinations of the original attributes. For many data sets, it is sufficient to consider only the first few PCs, thus reducing the dimension. However, for some data sets, PCA does not provide a satisfactory representation. It is not always the case that mutually-orthogonal linear combinations are the best way to define new attributes but e.g. nonlinear combinations needs to be sometimes considered. The analysis of the problem of dealing with data of high dimensionality is both difficult and subtle. The information loss caused by these methods is also sometimes a problem.

One of the latest methods in evolutionary computation is differential evolution (DE) algorithm [30]. In this paper we will examine the applicability of a classification method where data is first preprocessed with PCA and then the resulting data is classified with DE-classifier to the diagnosis of heart disease. In literature there are several papers where evolutionary computation research has concerned the theory and practice of classifier systems [4], [16], [17], [18], [31], [35], [10]. The differential evolution algorithm has been studied in unsupervised learning problems which can be in a sense repositioned to classification problem in [26], [11]. DE was also used combined with artificial neural networks in [1] for diagnosis of breast cancer. It is also been used to tune classifiers parameter values in [12] and in similarity classifier [23] to tune similarity measures parameters.